Generative Causal Explanations for Graph Neural Networks

Year: 2021 | Citations: 203 | Authors: Wanyu Lin, Hao Lan, Baochun Li

---

Abstract

This paper presents Gem, a model-agnostic approach for providing interpretable explanations for any GNNs on various graph learning tasks. Specifically, we formulate the problem of providing explanations for the decisions of GNNs as a causal learning task. Then we train a causal explanation model equipped with a loss function based on Granger causality. Different from existing explainers for GNNs, Gem explains GNNs on graph-structured data from a causal perspective. It has better generalization ability as it has no requirements on the internal structure of the GNNs or prior knowledge on the graph learning tasks. In addition, Gem, once trained, can be used to explain the target GNN very quickly. Our theoretical analysis shows that several recent explainers fall into a unified framework of additive feature attribution methods. Experimental results on synthetic and real-world datasets show that Gem achieves a relative increase of the explanation accuracy by up to 30% and speeds up the explanation process by up to $110\times$ as compared to its state-of-the-art alternatives.